Predicting Touchdowns
Final Report

By
Garrett JA Hass

1. Introduction

a. Problem Statement

NFL teams need help in assessing the Quarterback position. An important variable in a game is how many touchdowns they will throw. TDs have a huge impact on if they win or lose. The QB position is arguably the most important position on the field as they touch the ball on nearly every play. Their decisions can have a huge effect on the outcome of the game.

b. Criteria for Success

A team evaluates the performance of a QB to determine who to offer a contract, bench or start. Teams need to predict how many TDs a QB will throw in a game or a year. If a Quarterback is predicted to throw an excessive amount of touchdowns the organization could make a formable decision to offer a massive contract to this player. Pinpointing how many touchdowns a player will throw is our main objective. This will help the organization make a sound business decisions regarding the QB position.

2. Data

The database is the largest collection of quarterback data on Kaggle, (https://www.kaggle.com/datasets/speckledpingu/nfl-qb-stats). The stats are from every National Football League game from 1996-2016. This was a CSV file. The original columns were:

1. QB (Name)
2. Completion
3. Attempts
4. Yards
5. Yards per Attempt
6. Touchdowns
7. Interceptions
8. Rate
9. Long (Longest Throw)
10. Sack
11. Game Points
12. Loss (Yards from a Sack)
13. Home/Away

14. Year

I added:

1. Completion Percentage, (completions/attempts)
2. Yards per Completion
3. Touchdown per Completion
4. Touchdown per Attempt
5. Home
6. Away

The more variables in our model made it more accurate to predict how many touchdowns could be thrown. We also had to do some data cleaning as it was missing rows. The data had inaccurate entries such as -3 attempts in one game.

3. Data Wrangling and Cleaning

Initially we had 14 columns and 13,188 rows, (with some missing data) from the Kaggle data set. I did various cleaning methods on the data.

For example, I filled in a NaN value for the Rate column with the accurate Rate for Andrew Pinnock on row 8,644. I also had to drop the rows where it had more completions than attempts. This is physically impossible and probably was a simple data entry error. There were 97 rows that had an NaN value for Touchdown per Attempt simply because the player did not throw an attempt or a touchdown. I filled these with a 0. I ended up with 19 columns and 13,172 rows of information.

4. Exploratory Data Analysis

a. Correlations

We find various relationships within the data in the EDA phase. The strongest correlations were:

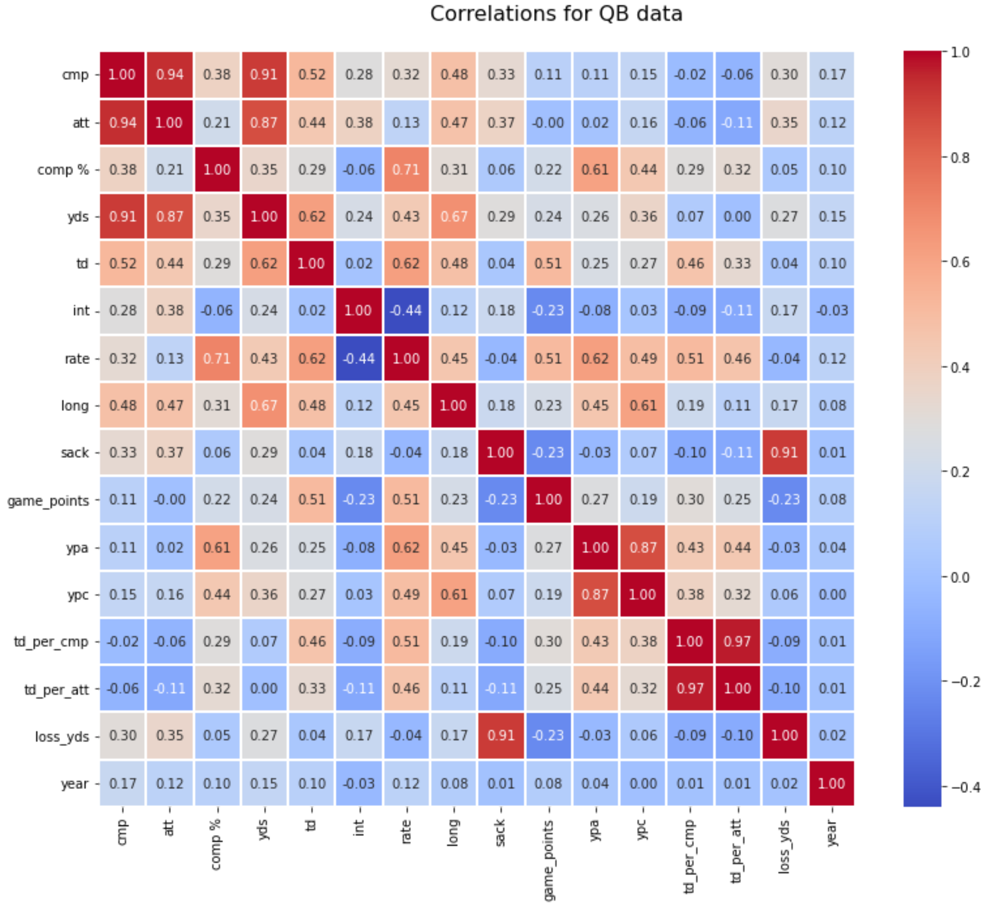
- Attempts & Completions (.94)
- Attempts & TDs (.87)
- Completions & Yards (.91)
- Attempts & yards (.87)
- Sack & loss_yds (.91)
- Yards per Catch & Yards per Attempt (.87)
- TD per Attempt & TD per completion (.97)

The largest negative correlations were:

- Game Points & Sack (-.23)
- Game Points & Loss yards (-.23)

- Interceptions & Rate (-.44)
- Game Points & Int (-.23).

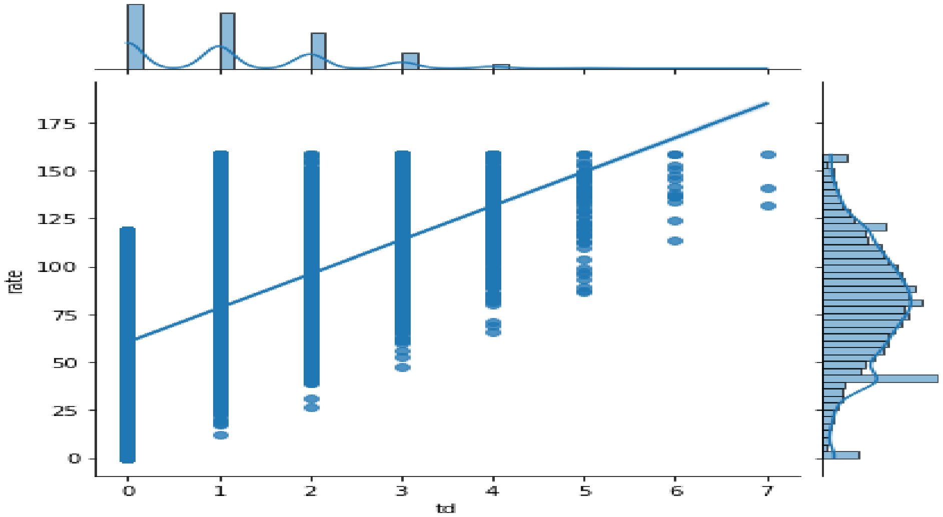
Three of these correlations included game points. Which would be very useful to a team as the more points you score the better odds you have of winning.



	cmp	att	comp %	yds	td	int	rate	long	sack	game_points	ypa	ypc	td_per_cmp	td_per_att	loss_yds	year
cmp	1.00	0.94	0.38	0.91	0.52	0.28	0.32	0.48	0.33	0.11	0.11	0.15	-0.02	-0.06	0.30	0.17
att	0.94	1.00	0.21	0.87	0.44	0.38	0.13	0.47	0.37	-0.00	0.02	0.16	-0.06	-0.11	0.35	0.12
comp %	0.38	0.21	1.00	0.35	0.29	-0.06	0.71	0.31	0.06	0.22	0.61	0.44	0.29	0.32	0.05	0.10
yds	0.91	0.87	0.35	1.00	0.62	0.24	0.43	0.67	0.29	0.24	0.26	0.36	0.07	0.00	0.27	0.15
td	0.52	0.44	0.29	0.62	1.00	0.02	0.62	0.48	0.04	0.51	0.25	0.27	0.46	0.33	0.04	0.10
int	0.28	0.38	-0.06	0.24	0.02	1.00	-0.44	0.12	0.18	-0.23	-0.08	0.03	-0.09	-0.11	0.17	-0.03
rate	0.32	0.13	0.71	0.43	0.62	-0.44	1.00	0.45	-0.04	0.51	0.62	0.49	0.51	0.46	-0.04	0.12
long	0.48	0.47	0.31	0.67	0.48	0.12	0.45	1.00	0.18	0.23	0.45	0.61	0.19	0.11	0.17	0.08
sack	0.33	0.37	0.06	0.29	0.04	0.18	-0.04	0.18	1.00	-0.23	-0.03	0.07	-0.10	-0.11	0.91	0.01
game_points	0.11	-0.00	0.22	0.24	0.51	-0.23	0.51	0.23	-0.23	1.00	0.27	0.19	0.30	0.25	-0.23	0.08
ypa	0.11	0.02	0.61	0.26	0.25	-0.08	0.62	0.45	-0.03	0.27	1.00	0.87	0.43	0.44	-0.03	0.04
ypc	0.15	0.16	0.44	0.36	0.27	0.03	0.49	0.61	0.07	0.19	0.87	1.00	0.38	0.32	0.06	0.00
td_per_cmp	-0.02	-0.06	0.29	0.07	0.46	-0.09	0.51	0.19	-0.10	0.30	0.43	0.38	1.00	0.97	-0.09	0.01
td_per_att	-0.06	-0.11	0.32	0.00	0.33	-0.11	0.46	0.11	-0.11	0.25	0.44	0.32	0.97	1.00	-0.10	0.01
loss_yds	0.30	0.35	0.05	0.27	0.04	0.17	-0.04	0.17	0.91	-0.23	-0.03	0.06	-0.09	-0.10	1.00	0.02
year	0.17	0.12	0.10	0.15	0.10	-0.03	0.12	0.08	0.01	0.08	0.04	0.00	0.01	0.01	0.02	1.00

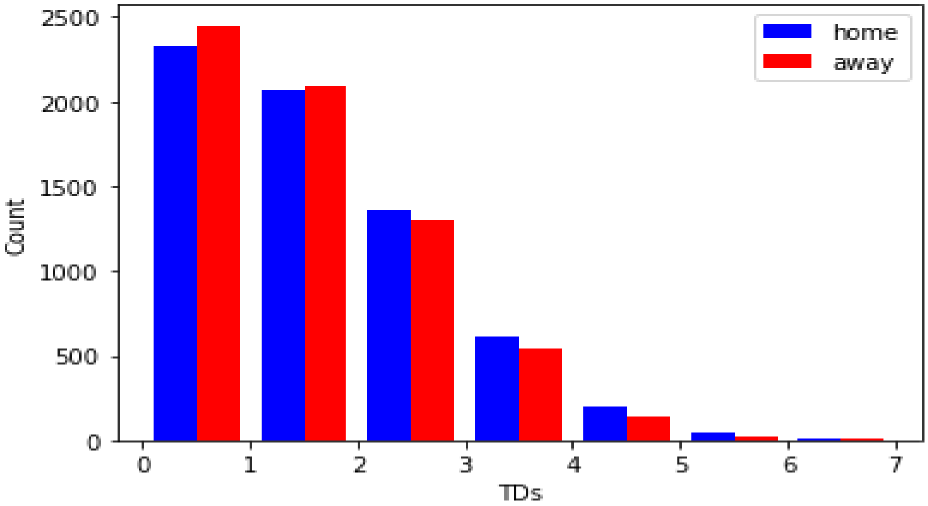
b. Touchdowns per Game

The distribution of the data showed us that many of the touchdowns thrown were in the range of 0-3 and that there were not many 4-7 touchdowns thrown in a game. Very hard to do.There is a wide range of rates when a QB throws 0 TDs and a small range of rates when they throw 6-7 TD. A player can have a high rate and not throw many TDs.



c. Home and Away Touchdowns

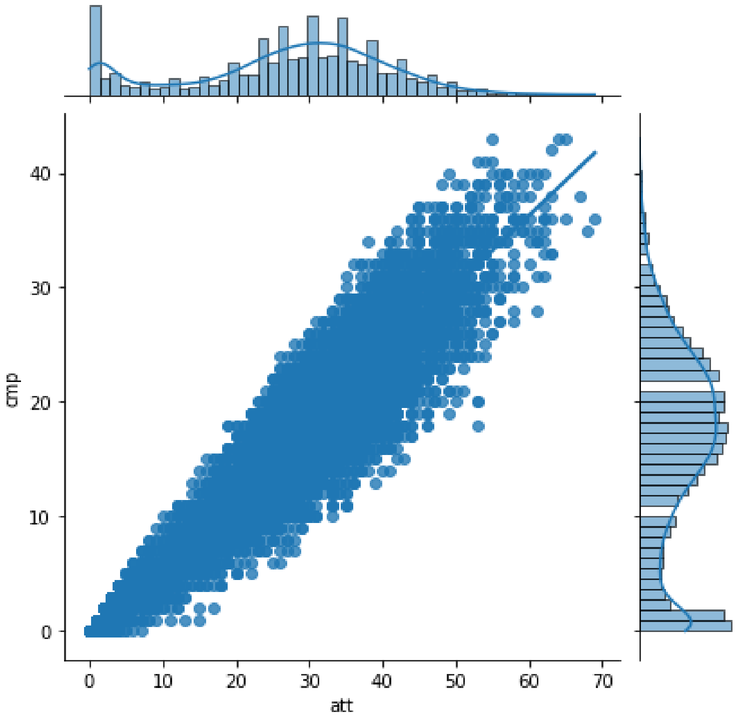
I was interested to see if a QB is more prone to throw more touchdowns at their home stadium than away. There is less noise when calling an audible on the line of scrimmage. Also players are more comfortable when they can drive to the stadium as opposed to traveling and staying in a hotel for an away game.



Evidently that is not the case. Home and away touchdowns thrown per game are very relevant. They are very close no matter where the location of the game is.

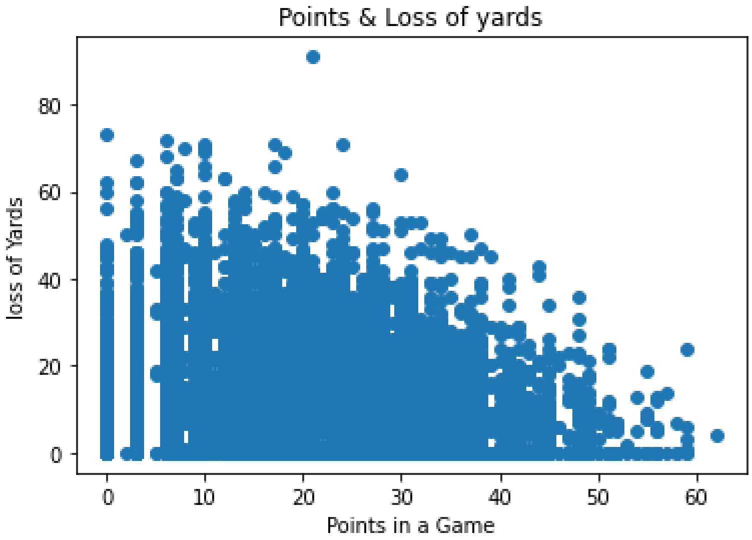
d. Completions & Attempts

These two variables go hand in hand. I was curious to see if they had a positive correlation. Meaning as one goes up the other variable ascends as well.



If the player completed more passes there is a greater likelihood that they threw more touchdowns. Which could be beneficial for the team to win the game. Here we can see that it was a positive correlation of .91.A coach will call more pass plays if the QB is completing a large amount of passes. This leads to a greater amount of touchdowns. Potentially more attempts lead to more touchdowns if the QB is completing those passes.

e. Points Scored and Loss of Yards



An interesting finding was the negative correlation between both of these variables. As the amount of loss yards increased the points per game for a team decreased. More loss yards from a sack led to less opportunities for a team to score points. If the offensive line did a good job of reducing sacks, the team scored more points and likely won the game. Having a good offensive line can be beneficial to winning a game.

5. Modeling Phase

Initially I did a train test split on the data to see how accurately each of the four models I examined performed on the data. The four models I tried were linear regression, lasso regression, ridge regression and random forest regression. The best performing model was random forest regression.

The Random Forest R2 score was nearly perfect at .997. A very solid MAE of .006 and RMSE of .06 on the test data. The CV average & Grid Search CV for Random Forest was a very high score of .9977 and .9979.

These scores were much higher than the other regression models. The other three had an R2 score in the upper .70s. The RMSE scores for linear, ridge and lasso were all around .378 as well.

6. Recommendations

After exploring our model and data I recommend these three ideas:

a. Free Agency

Going forward I suggest we evaluate potential free agent QB statistics based on the random forest model. We could run their historical numbers in our model to predict how many TDs they would throw in a game or year. If they get sacked a lot regardless of the offensive line they might not score as many points for the team. We could make a sound decision on whom to spend millions of dollars on in free agency. If we choose the right QB it could change the future of the franchise immediately.

b. In-game decisions

We could use this model on which QB to bench or start during the game. If a player is having a bad quarter or half we could bench them based upon how many touchdowns they could throw in the next portion of the game. For the other team we could predict how many touchdowns they might throw based upon their stats in the first half and adjust our defensive scheme accordingly.

c. Points per Game

We could also do a train test split and find a good model on how many points we score in a game. Given the other QB stats we could predict how many points we will score in the game. If we play a team that has a very good offense we could show the stats to our QB so they know what kind of game they need to play to score that many points to win. Do they need to throw for many yards or not turn over the ball to score 32 points for that particular game? These are the types of questions we can answer with this model. I would also explore the negative correlation we found in the heat map between sack, loss yards and interceptions with game points. This could give us an advantage on how to score more points.

7. Further Research

Metrics in the NFL is an interesting concept. There are many variables that we cannot quantify, like a player’s will and determination at any given moment. The future is hard to predict. This random forest model to predict touchdowns a QB will throw is very helpful in understanding what could potentially happen in a game. It gives us a chance to comprehend how a player might do. The more information the better we can make decisions.

I would run models on every stat we have on defense, special teams and offense to predict how every player might perform at each position. This would help to evaluate free agents and our own players. We could improve on who to sign in the off-season. This could potentially lead to more wins and revenue for the organization.